AUTOMATED MORPHOLOGICAL ANALYZER GENERATOR OF ANGIKA

A

SYNOPSIS

Submitted in partial fulfillment of the

requirements for the award of the degree

of

Doctor of Philosophy

in the

Department of

COMPUTER SCIENCE & ENGINEERING

Submitted to

YBN UNIVERSITY

RAJAULATU, NAMKUM, RANCHI

JHARKHAND – 834010, (INDIA)

Jan, 2023

Research Supervisor:
Dr. Manisha Kumari Deep
Professor, School of Computer Science & IT
YBN University, Ranchi

Research Scholar:
Alok Kumar
Department of Computer Science & Engineering
YBN University, Ranchi
Enrollment No.: YBNU2003277

Abstract

Angika, an Eastern Indo-Aryan language spoken by over seven million people across Bihar, Jharkhand, and adjoining regions, possesses a morphologically complex grammatical system yet remains entirely absent from computational linguistic research. The absence of foundational NLP tools — particularly a morphological analyzer — prevents the development of any higher-level language technology for the language.

This research designs and develops an Automated Morphological Analyzer Generator for Angika using a Finite-State Transducer (FST) framework implemented through HFST, augmented by a hybrid rule-based module that handles morphophonemic alternations and paradigmatic irregularities beyond the scope of pure finite-state representations. A generator-based architecture is proposed wherein a functional analyzer is automatically compiled from a structured lexical database and formally specified morphological rules in lexc/twolc notation, keeping linguistic knowledge decoupled from processing logic.

The system will be evaluated using Precision, Recall, and F1-score on a manually verified gold-standard dataset. The research delivers the first computationally formalized account of Angika morphology and establishes a reusable NLP framework applicable to other low-resource Indian languages.

Keywords: Angika, Morphological Analysis, FST, HFST, Hybrid Rule-Based NLP, Low-Resource Languages, Computational Morphology

1. Introduction

Natural Language Processing (NLP) constitutes a multidisciplinary domain at the intersection of computational science and linguistics, with its foundational goal being the development of systems capable of analyzing, interpreting, and generating human language automatically. Among the various sub-tasks encompassed within NLP, morphological analysis occupies a particularly central position, as it provides the structural basis upon which all higher-level linguistic processing depends.

Morphological analysis refers to the computational process of decomposing surface word forms into their constituent morphemes — roots, stems, and affixes — while simultaneously identifying associated grammatical attributes such as tense, aspect, mood, number, gender, person, case, and honorificity. For morphologically rich languages, where a single lexical root may produce dozens or even hundreds of surface realizations through systematic inflectional and derivational processes, the availability of a robust morphological analyzer is a prerequisite for any meaningful NLP application.

The Indian subcontinent hosts a vast array of morphologically complex languages belonging to the Indo-Aryan, Dravidian, Tibeto-Burman, and Austro-Asiatic families. While high-resource languages such as Hindi, Bengali, and Tamil have witnessed considerable progress in NLP tool development, a substantial number of regional and minority languages continue to remain outside the scope of computational linguistic research. This asymmetry perpetuates digital marginalization for speakers of these languages.

Angika is an Eastern Indo-Aryan language spoken predominantly in the Anga region of Bihar and Jharkhand, with significant speaker communities in parts of West Bengal and the Terai belt of Nepal. Classified under ISO 639-3 code anp, the language is estimated to have over seven million speakers. Despite possessing a well-structured and rule-governed grammatical system, Angika remains severely underrepresented in both descriptive linguistics and computational language research.

The morphological architecture of Angika is characterized by extensive inflectional paradigms operating across nominal, verbal, adjectival, and postpositional categories. Verbal morphology in particular exhibits intricate layers of agreement governed by tense, aspect, mood, person, number, and honorificity — a complexity that makes automated processing both challenging and linguistically significant.

The present research proposes the design and development of an Automated Morphological Analyzer Generator for Angika, employing a Finite-State Transducer (FST) based framework augmented with a hybrid rule-based strategy. This approach combines the computational elegance and efficiency of finite-state methods with the linguistic expressivity required to handle the exceptional morphophonemic and paradigmatic complexity of Angika — without reliance on large annotated corpora that are currently unavailable for the language.

2. Need of the Study

The justification for undertaking the present research emerges from a convergence of linguistic, computational, and socio-cultural imperatives that collectively underscore the urgency of developing morphological processing tools for Angika.

At the computational level, Angika is entirely absent from the inventory of languages supported by existing NLP platforms, corpora repositories, and language technology frameworks. No morphological analyzer, part-of-speech tagger, syntactic parser, or machine translation system currently exists for the language. This void effectively excludes Angika from all modern digital applications, including search engines, spell-checkers, virtual assistants, and translation systems.

At the linguistic level, Angika displays a morphological richness that demands dedicated computational modeling. Its verbal system, in particular, encodes multiple interacting grammatical categories simultaneously, generating surface forms whose analysis cannot be approximated through tools developed for structurally different languages such as Hindi. Direct adaptation of Hindi NLP tools to Angika is neither technically feasible nor linguistically valid due to fundamental differences in morphological paradigms, agreement structures, and phonological alternation patterns.

From a methodological standpoint, data-driven and neural NLP approaches — which have achieved remarkable results for high-resource languages — are fundamentally unsuitable for Angika at the present stage, owing to the complete absence of large-scale annotated training corpora. Rule-based finite-state approaches offer the most viable and principled alternative, providing linguistic interpretability, computational efficiency, and the ability to function effectively with minimal data resources.

Beyond academic motivations, the development of morphological processing tools for Angika carries direct socio-linguistic significance. Computational tools are essential for the digital documentation, preservation, and revitalization of endangered and low-resource languages. By enabling Angika's integration into digital platforms and educational technologies, the present research makes a concrete contribution to linguistic diversity and the empowerment of Angika-speaking communities.

3. Problem Statement

Despite the structural regularity and morphological complexity that characterize Angika, the language currently possesses no automated system capable of performing morphological analysis. This absence constitutes a foundational gap that prevents the development of any higher-level NLP application for the language.

Angika word forms undergo layered morphological transformations governed by grammatical parameters including number, gender, person, tense, aspect, mood, case, and honorificity. A single verbal root, for instance, may yield a broad range of distinct surface realizations depending on the configuration of these parameters and the application of associated morphophonemic rules. These processes are well-

documented at a descriptive level but have never been formalized in a computationally tractable representation.

The non-existence of a morphological analyzer for Angika creates several cascading deficiencies. There exists no mechanism for automated root identification, morpheme segmentation, or grammatical feature extraction from surface forms. Consequently, the development of downstream NLP tools — including taggers, parsers, information retrieval systems, and translation engines — remains impossible. Furthermore, the inapplicability of data-driven methods due to corpus absence means that a rule-governed computational approach is the only feasible path forward.

An additional challenge arises from the generator-level gap: not only is a working morphological analyzer absent, but there also exists no systematic framework for generating such an analyzer from a defined linguistic knowledge base. The proposed research addresses this dual gap by constructing an Automated Morphological Analyzer Generator — a system that produces a functional analyzer from an organized lexicon and a formally specified rule set, thereby ensuring reusability, extensibility, and linguistic transparency.

4. Research Hypothesis

The present research is guided by the following primary hypothesis:

A Finite-State Transducer (FST) based morphological analyzer, when augmented with a hybrid rule-based component addressing morphophonemic alternations and paradigmatic exceptions, can effectively and accurately analyze the inflectional and derivational complexity of Angika word forms — without requiring large annotated corpora — thereby producing a reliable, interpretable, and extensible computational tool suitable for a low-resource language environment.

This hypothesis further implies that:

- Angika morphological processes are sufficiently regular and rule-governed to be formally captured through finite-state representations.
- A generator-based architecture — separating linguistic knowledge from processing logic — will yield a more adaptable and maintainable system than a static analyzer.
- The hybrid integration of FST with explicit rule-based handling of exceptions will outperform a purely transducer-based system in terms of coverage and accuracy on Angika word forms.

5. Objectives of the Research

The overarching objective of this research is to design, develop, and evaluate an Automated Morphological Analyzer Generator for Angika that is linguistically principled, computationally efficient,

and viable within a low-resource language setting. This primary objective is operationalized through the following specific aims:

1. To conduct a systematic and exhaustive linguistic analysis of Angika morphology, encompassing all major lexical categories — nouns, verbs, adjectives, and postpositions — with respect to their inflectional and derivational behavior.
2. To identify, classify, and formally document the inflectional paradigms, derivational processes, and morphophonemic alternation patterns operative in Angika, with specific attention to grammatical categories of number, gender, person, tense, aspect, mood, case, and honorificity.
3. To represent identified morphological processes through formally specified linguistic rules amenable to finite-state computational modeling.
4. To develop a Finite-State Transducer (FST) based morphological analyzer for Angika, implementing the formalized rule system within an HFST (Helsinki Finite-State Toolkit) framework.
5. To design a hybrid rule-based component that complements the FST layer by handling morphophonemic alternations, paradigmatic irregularities, and boundary phenomena not easily captured through finite-state transduction alone.
6. To construct a morphological analyzer generator — a meta-level system capable of producing a functional analyzer from a specified Angika lexical database and rule set — ensuring modularity, reusability, and ease of future extension.
7. To compile a dedicated Angika lexical database comprising root forms, grammatical feature specifications, and morpheme inventories required for morphological processing.
8. To evaluate the proposed system's performance using manually verified linguistic data, measuring accuracy in root identification, morpheme segmentation, and grammatical feature assignment through standard metrics of Precision, Recall, and F1-score.
9. To propose a reusable computational framework applicable to other low-resource Indian languages exhibiting comparable morphological characteristics.

6. Research Questions

The present study is organized around the following central research questions:

1. What are the principal morphological processes — inflectional, derivational, and morphophonemic — that govern word formation across major lexical categories in Angika?
2. How can the morphological paradigms of Angika be formally represented in a computationally tractable manner using finite-state formalisms?
3. To what degree can a Finite-State Transducer architecture, augmented by a hybrid rule-based component, model the full scope of Angika's morphological complexity in the absence of annotated corpora?

4. What architectural design principles best support a morphological analyzer generator that remains extensible, linguistically transparent, and reusable across related low-resource languages?
5. How accurately does the proposed hybrid FST-based system analyze Angika surface word forms with respect to root identification, morpheme segmentation, and grammatical feature assignment?
6. What broader contribution does the developed system make toward enabling higher-level NLP applications and the digital preservation of Angika?

7. Review of Related Literature

The present research is situated within a rich body of scholarly work spanning descriptive morphological linguistics, finite-state computational methods, and NLP research on Indian and low-resource languages. A selective but representative review of this literature is presented below.

7.1 Theoretical Foundations of Morphology

The foundational theoretical framework for morphological analysis derives from structuralist and generative traditions. Hockett (1954) established the conceptual distinction between item-and-arrangement and item-and-process models of morphology, both of which inform subsequent computational implementations. Aronoff (1976) formalized word-formation rules within a generative framework, while Anderson (1992) elaborated the A-morphous morphology model foregrounding the role of paradigmatic relations — a perspective directly relevant to Angika's complex verbal agreement system.

7.2 Finite-State Methods in Morphological Analysis

The application of finite-state methods to morphological analysis was systematically developed by Koskenniemi (1983), whose two-level morphology model laid the computational groundwork for all subsequent FST-based analyzers. Beesley and Karttunen (2003) extended this foundation with the Xerox Finite-State Toolkit, demonstrating the scalability of FST approaches across typologically diverse languages. The Helsinki Finite-State Toolkit (HFST), developed by Linden et al. (2011), provides an open-source, production-grade implementation of the same paradigm and constitutes the primary computational framework adopted in the present research. Antworth (1990) further consolidated the methodology of PC-Kimmo and lexical phonology within finite-state computational systems.

7.3 Morphological Analysis for Indian Languages

Computational morphology for Indian languages has been addressed with varying degrees of depth. Bharati et al. (1995) developed a comprehensive morphological analysis framework for Hindi, while Goyal and Lehal (2008) produced a Punjabi morphological analyzer using rule-based approaches. Dash and Pattnaik (2005) addressed Odia morphology from a computational perspective. More recently, Chakrabarty et al. (2017) proposed a data-light hybrid approach for Bengali morphological

disambiguation. These works collectively demonstrate the viability of rule-based and hybrid strategies for morphologically rich languages with limited annotated resources — a context directly analogous to that of Angika.

7.4 NLP Research on Low-Resource and Minority Languages

The challenge of developing NLP tools for low-resource languages has received growing attention in the computational linguistics community. Bird (2009) articulated a framework for endangered language documentation emphasizing computational tool development. Cieri et al. (2016) reviewed resource creation strategies for under-resourced languages, advocating for rule-based bootstrapping methods in data-scarce environments. Eryigit and Adalı (2004) demonstrated the efficacy of finite-state approaches for Turkish — a morphologically rich, low-resource language — providing a methodologically analogous precedent for the present work.

7.5 Descriptive Linguistics of Angika

Descriptive linguistic work on Angika remains sparse. Jha (1979) provides one of the earliest systematic accounts of Angika grammar, including partial documentation of its nominal and verbal morphology. Kumar (2008) offers a comparative treatment of selected morphological features within the broader Maithili-Angika continuum. Grierson's Linguistic Survey of India (1903) contains early observations on Anga dialect morphology, though the descriptions are broad and lack formal precision. The present research builds directly upon these descriptive accounts, translating their insights into computationally tractable formal representations for the first time.

8. Scope of the Study

The scope of the present research is defined along both linguistic and computational dimensions, with a deliberate focus on depth and precision over broad coverage.

Linguistically, the study encompasses the morphological analysis of four major lexical categories in Angika: nouns, verbs, adjectives, and postpositions. Within each category, the research examines inflectional processes associated with relevant grammatical features — number and case for nominals; tense, aspect, mood, person, number, and honorificity for verbs — as well as derivational processes that generate new lexical items from existing roots. Morphophonemic alternations at morpheme boundaries receive systematic treatment across all categories.

Computationally, the scope encompasses the formal specification of morphological rules, the construction of a finite-state transducer network using HFST, the development of a hybrid rule-based supplementary module, and the design of an analyzer generator architecture that encapsulates linguistic knowledge

separately from processing logic. The scope further includes the compilation of an Angika lexical database and the evaluation of the complete system on a manually verified test dataset.

The study deliberately excludes the development of large-scale annotated corpora, machine-learning-based models, syntactic parsing systems, and machine translation engines. These represent valuable areas for future research but fall outside the focused scope of the present investigation.

9. Research Methodology

The methodology of the present research integrates descriptive linguistic fieldwork with formal computational modeling, organized into six interrelated phases. The choice of an FST-based hybrid approach is both principled and pragmatic — principled in its linguistic interpretability, and pragmatic in its suitability for a low-resource language environment.

9.1 Linguistic Data Collection

The initial phase involves systematic collection of Angika linguistic data from multiple complementary sources: structured elicitation sessions with native speakers, analysis of available written texts in Angika, and review of existing descriptive grammatical accounts. Data collection targets representative word forms across all morphological categories, with particular emphasis on paradigm completeness for verbal forms, which exhibit the highest degree of morphological complexity in the language. All collected data undergoes native-speaker verification to ensure authenticity.

9.2 Morphological Classification and Feature Specification

In the second phase, collected word forms are systematically classified into lexical categories and analyzed with respect to their associated morphological features. Nouns are analyzed for number, gender, case, and postpositional attachment patterns. Verbs are analyzed for tense, aspect, mood, person, number, and honorificity. Adjectives are examined for agreement with nominal heads, and postpositions are analyzed for case-marking behavior. Feature matrices are constructed for each paradigm to guide subsequent rule formalization.

9.3 Rule Formalization for FST

Based on the linguistic analysis, morphological rules are formally specified as finite-state transduction patterns. Each rule defines a mapping between an underlying lexical representation and a surface phonological form, specifying valid root-affix combinations, morpheme-ordering constraints, and conditions governing morphophonemic alternations. Rules are organized modularly by lexical category and grammatical feature, facilitating systematic testing and incremental refinement. The formalization conforms to the two-level morphology framework as implemented in HFST.

9.4 FST-Based Hybrid Architecture

The core computational system combines two complementary processing layers. The primary layer consists of a Finite-State Transducer network compiled from the formalized morphological rules using the Helsinki Finite-State Toolkit (HFST). FST networks handle the predominant class of regular, concatenative morphological processes with high efficiency. The secondary layer is a hybrid rule-based module designed to address morphophonemic alternations and paradigmatic irregularities whose conditional complexity exceeds the expressive capacity of pure FST representations. This module applies post-FST filtering and correction through explicit linguistic rules encoded in Python-based processing scripts. The two layers interact through a structured interface that maintains the modularity of the overall system.

9.5 Analyzer Generator Design

A key architectural contribution of the present research is the design of a morphological analyzer generator — a meta-level system that produces a functional morphological analyzer from a specified linguistic knowledge base. Unlike a static analyzer hardcoded for a fixed dataset, the generator accepts as input an Angika lexical database (containing root forms and grammatical feature specifications) and a rule set (formalized in HFST-compatible lexc/twolc notation), and automatically compiles these into a working analyzer. This design ensures that the linguistic knowledge component remains fully decoupled from the processing engine, enabling straightforward modification, extension, and reuse without restructuring the system's core architecture.

9.6 Lexical Database Construction

A dedicated Angika lexical database is constructed to serve as the knowledge source for the analyzer generator. The database comprises root forms across all covered lexical categories, along with associated grammatical feature specifications, morphological class assignments, and frequency annotations where available. Database entries are organized in a structured format compatible with HFST lexc notation, ensuring direct usability within the computational pipeline.

9.7 Evaluation Methodology

The evaluation of the proposed system is conducted on a manually constructed gold-standard test dataset comprising verified Angika word forms spanning all covered lexical categories and morphological phenomena. System performance is measured using standard information retrieval metrics: Precision (proportion of system-generated analyses that are correct), Recall (proportion of correct analyses successfully retrieved by the system), and F1-score (harmonic mean of Precision and Recall). Separate evaluations are conducted for root identification accuracy, morpheme segmentation accuracy, and grammatical feature assignment accuracy, enabling a granular assessment of system strengths and limitations.

10. Tools and Technologies

The following software tools and technologies constitute the computational infrastructure of the present research:

Tool / Framework	Purpose	Justification
HFST (Helsinki Finite-State Toolkit)	FST compilation, lexicon encoding (lexc), two-level rule specification (twolc)	Open-source, widely used for morphological analysis of low-resource languages
Python 3.x	Hybrid rule-based module, pre/post-processing, evaluation scripts	Flexible scripting for rule application and data pipeline management
NLTK (Python)	Tokenization, test data preprocessing, corpus utilities	Standard NLP toolkit with active community support
lexc / twolc notation	Morphological rule specification compatible with HFST	Declarative, linguistically interpretable rule formalism
Devanagari Unicode	Script encoding for all Angika data and lexical resources	Standard script used for Angika written representation
XML / JSON	Lexical database storage and interchange format	Structured, portable format for linguistic feature data

11. Expected Outcomes of the Study

The present research is anticipated to yield the following substantive contributions to the fields of computational linguistics and NLP for low-resource languages:

- A comprehensive and formally documented morphological description of Angika, covering inflectional, derivational, and morphophonemic phenomena across all major lexical categories — constituting the most detailed computational linguistic account of Angika morphology to date.
- A formally specified, computationally implementable rule set capturing the morphological structure of Angika in HFST-compatible lexc and twolc notation.
- A functional Finite-State Transducer (FST) network for Angika morphological analysis, compiled using HFST from the formalized rule set and lexical database.
- A hybrid rule-based module that augments the FST layer by handling morphophonemic alternations and paradigmatic irregularities beyond the expressive scope of pure finite-state representations.

- An Automated Morphological Analyzer Generator for Angika — a meta-level system producing a functional analyzer from a defined linguistic knowledge base, enabling modular extension and reuse.
- A dedicated Angika lexical database comprising root forms, morphological class annotations, and grammatical feature specifications.
- A gold-standard evaluation dataset of verified Angika word forms, along with a systematic performance evaluation reporting Precision, Recall, and F1-score metrics.
- A reusable and extensible computational framework applicable to other low-resource Indian languages with analogous morphological characteristics.

12. Significance and Original Contribution

The significance of the present research is multidimensional, extending across the domains of theoretical linguistics, applied NLP, and socio-linguistic practice.

From a linguistic standpoint, the study provides the first formally specified, computationally oriented account of Angika morphology. By translating descriptive linguistic observations into precise computational representations, the research bridges a critical gap between traditional grammatical scholarship and modern language technology.

From a computational standpoint, the research demonstrates the viability and effectiveness of an FST-based hybrid architecture for a morphologically rich language in a zero-corpus environment. The generator-based design — which separates the linguistic knowledge base from the processing engine — represents an architectural contribution that extends beyond Angika to inform tool development for other low-resource languages.

The originality of the present work resides primarily in its hybrid generator architecture. Unlike static morphological analyzers designed for specific language snapshots, the proposed system generates a functional analyzer from an updatable linguistic knowledge base. This design supports iterative refinement, community-driven extension, and adaptation to related languages — capabilities that are particularly valuable in the low-resource language context where resources are perpetually evolving.

From a socio-cultural standpoint, the research contributes directly to the digital inclusion and technological empowerment of Angika-speaking communities by making the language processable in computational systems, thereby enabling its integration into educational technologies, digital media platforms, and governmental digital services.

13. Tentative Chapter Scheme

Chapter	Title	Key Content
1	**Introduction**	Research background, Angika profile (speakers, ISO code, script), role of morphology in NLP, objectives, scope, and chapter overview
2	**Review of Literature**	FST methods, Indian language NLP, Angika descriptive linguistics, low-resource NLP strategies, gap identification
3	**Morphological Structure of Angika**	Detailed linguistic analysis of nouns, verbs, adjectives, postpositions; inflectional paradigms; derivational processes; morphophonemic alternations
4	**Formalization of Morphological Rules**	Finite-state rule specification in lexc/twolc notation; feature structures; morphophonemic constraint encoding
5	**System Architecture: FST**	FST network design using HFST; hybrid rule-based module; generator architecture; lexical database structure and format
6	**System Architecture: Hybrid Design**	
7	**Implementation and Evaluation**	Computational implementation details; gold-standard test dataset construction; Precision/Recall/F1 evaluation; error analysis
8	**Conclusion and Future Directions**	Summary of findings, contributions, limitations, proposed future extensions to related languages and higher-level NLP tasks

14. Research Timeline

The proposed research is planned for completion within the standard prescribed duration of the PhD programme, organized as follows:

Duration	Phase	Activities
Months 1–6	Literature Review & Framework Design	Comprehensive review of FST methods, Indian language NLP, Angika descriptive literature; refinement of research framework and methodology
Months 7–12	Linguistic Data Collection & Analysis	Native speaker elicitation, written text analysis, morphological paradigm documentation, feature matrix construction
Months 13–24	Rule Formalization & FST Development	Formal rule specification in lexc/twolc, HFST network compilation, hybrid module design, lexical database construction
Months 25–30	Analyzer Generator & System Integration	Generator architecture implementation, integration of FST and hybrid layers, system testing and debugging
Months 31–36	Evaluation & Refinement	Gold-standard dataset construction, Precision/Recall/F1 evaluation, error analysis, system refinement based on results

Months 37–42	Thesis Writing & Submission	Drafting, revision, formatting, and submission of the complete doctoral thesis

15. References

Anderson, S. R. (1992). A-Morphous Morphology. Cambridge University Press.

Antworth, E. L. (1990). PC-KIMMO: A Two-Level Processor for Morphological Analysis. Summer Institute of Linguistics.

Aronoff, M. (1976). Word Formation in Generative Grammar. MIT Press.

Beesley, K. R., & Karttunen, L. (2003). Finite State Morphology. CSLI Publications, Stanford.

Bharati, A., Chaitanya, V., & Sangal, R. (1995). Natural Language Processing: A Paninian Perspective. Prentice-Hall of India.

Bird, S. (2009). Natural language processing and linguistic fieldwork. Computational Linguistics, 35(3), 469–474.

Chakrabarty, A., Mondal, S., & Bandyopadhyay, S. (2017). Hybrid approach for Bengali morphological disambiguation. Proceedings of ICON 2017.

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016). Selection criteria for low resource language programs. Proceedings of LREC 2016.

Dash, N. S., & Pattnaik, P. K. (2005). Computational morphology of Odia: A rule-based approach. In Proceedings of NLPAI 2005.

Eryigit, G., & Adalı, E. (2004). An affix stripping morphological analyzer for Turkish. Proceedings of IASTED NLP 2004.

Goyal, V., & Lehal, G. S. (2008). Hindi morphological analyzer and generator. Proceedings of ICETET 2008.

Grierson, G. A. (1903). Linguistic Survey of India, Vol. V: Indo-Aryan Family. Government of India Press, Calcutta.

Hockett, C. F. (1954). Two models of grammatical description. Word, 10(2–3), 210–231.

Jha, G. N. (1979). A Grammar of Angika. Doctoral Dissertation, Banaras Hindu University.

Koskenniemi, K. (1983). Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. University of Helsinki.

Kumar, A. (2008). Morphological features of Angika in comparative perspective. Journal of Bihar Studies in Language, 4(1), 22–37.

Linden, K., Silfverberg, M., Pirinen, T., & Ruokolainen, T. (2011). HFST — Framework for compiling and applying morphologies. Proceedings of the 2nd International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011), Lecture Notes in Computer Science, Vol. 100.